Performance Evaluation of a Queue Fed by a Poisson Pareto Burst Process

Ronald G. Addie*
University of Southern Queensland
Toowoomba Qld 4350, Australia
Phone: (+61 7) 4631 5520
Fax: (+61 7) 4631 5550
Email: addie@usq.edu.au

Timothy D. Neame and Moshe Zukerman
ARC Special Research Centre for Ultra-Broadband Information Networks
Department of Electrical and Electronic Engineering
The University of Melbourne
Victoria, 3010, Australia.
Phone: (+61 3) 8344 3809
Fax: (+61 3) 8344 3823
Email: {t.neame, m.zukerman}@ee.mu.oz.au

Abstract

This paper provides means for performance evaluation of a queue with Poisson Pareto Burst Process (PPBP) input. Because of the long range dependent nature of the PPBP, straightforward simulations are unreliable. New analytical and simulation techniques are described in this paper. Numerical comparison between the results shows consistency. Conservative dimensioning rules using zero buffer approximations are examined versus the more aggressive analytical approach based on the results of this paper to provide practical guidelines for network design.

1 Introduction

The Poisson Pareto Burst Process (PPBP), also known as the M/Pareto process [4, 16, 17, 18], is a specific form of the Poisson Burst Process [22] or the M/G/∞ process [11, 20, 25, 29]. The PPBP has gained its appeal because it is formed in a way consistent with the way people generate Internet traffic, and because it can be used to model real traffic streams.

The PPBP also has the highly attractive property that its variance-time curve (the variance of the total traffic arriving in an interval of length t, as a function of t) is asymptotically, for large t, the same as fractional Brownian noise with Hurst parameter $H > 0.5$, which is the form that has been observed in real traffic in many studies [13]. In fact, the variance-time curve (and therefore also the autocovariance) of the PPBP can be made as close as required to that of any long-range dependent fractional Brownian noise process by appropriate choice of parameters.

*Corresponding author

The PPBP is based on an underlying Poisson process representing points in time at which any of a large number of users begins transmitting a traffic burst. It has been shown [7] that the distribution of the sizes of files transmitted across the Internet is heavy-tailed. If we assume that each file is transferred in a single burst, it seems reasonable to model the burst lengths by a heavy-tailed random variable. The Pareto distribution has been chosen to model the heavy-tailed burst lengths. In this paper, we choose the parameters of the Pareto distribution such that the burst length will have infinite variance but finite mean.

Given its practical viability, the PPBP and its associated queueing problems have been extensively studied. Tsybakov and Georganas have been amongst the leaders in studies of these processes. Together with Likhanov they showed in [14] that a model such as the PPBP could be considered a limiting case for the multiplexing of a large number of independent on-off sources with heavy tailed on and/or off time distributions. They examined the conditions under which M/G/∞ processes are self-similar in [27]. Results regarding the properties of burst processes in which the transmission rate of each burst is not necessarily constant, or the same for all bursts, are presented in [28].

A number of authors have also worked to develop approximations for the stationary queueing distribution of the PPBP and related processes. By far the most common approach to developing approximations for stationary queueing distributions is to develop a formula which is exact, or which provides upper and lower bounds, for a limiting case, as a certain parameter tends to a specific value. Typically the approximation is assumed to provide a satisfactory approximation for values of this parameter which are sufficiently close to the limiting value.

Large deviations theory has provided a fruitful approach, and in the present context it has been used to provide asymptotically accurate upper and lower bounds in the usual large deviations sense as $x \to \infty$, where x is the buffer contents. Examples of this approach may be found in the work of Tsoukatos and Makowski [25], Parulekar and Makowski [20], González-Arévelo and Samorodnitsky [9] and Mandjes [15]. In [29], Tsybakov and Georganas use a large deviations approach to find expressions for both upper and lower bounds of the queueing performance of a finite buffer single server queue (SSQ) fed by an M/G/∞ process.

A different limiting case is considered by Addie, Mannersalo and Norros in [1]. They consider an approximation which is accurate for a Gaussian traffic stream. The queueing performance of the PPBP tends towards Gaussian as λ, the level of multiplexing in the stream, increases [4]. Thus the approximation in [1] is accurate in the limit as $\lambda \to \infty$.

Despite these efforts, accurate performance results for an SSQ fed by a PPBP under normal traffic conditions have not been achieved. Although the approximations are good *in the limit*, they are not satisfactory for values of real interest, which happen to be not sufficiently close to the limiting value for the quality of the approximation to be satisfactory. The difficulty of estimating the queueing performance of the PPBP is compounded by the fact that the infinite variance of the burst size means that straightforward simulations of PPBP SSQs are unreliable.

In this paper, we present new insight into the PPBP queue behaviour, based on the idea of dividing the process into a pair of independent sub-processes. One of these sub-processes will be made up of *short bursts*, while the other sub-process is made up of *long bursts* which will last longer than the time scale of interest. We use this insight to develop new techniques for dealing with the complexities of the PPBP.

The contribution of this paper is fourfold: (1) we provide a new analytical method for obtaining the queueing performance of a PPBP SSQ; (2) we provide a technique to improve the reliability of PPBP SSQ simulation; (3) we use the improved simulations to show that the analytical method is accurate; and (4) we provide dimensioning guidelines considering the above in conjunction with the zero buffer approximation (ZBA).

The remainder of the paper is organized as follows. In Section 2, we provide a formal description of the PPBP and discuss its peculiarities related to long bursts. In Section 3, we motivate and present our improved simulation technique for a PPBP SSQ. In Section 4, the analytical approximation is obtained. Then, in Section 5, we present numerical results and discuss the accuracy of the different approaches. Finally, in Section 6, we discuss queue

dimensioning implications.

2 The Poisson Pareto Burst Process

As mentioned in the introduction, the PPBP is attractive as a model for Internet traffic because it has some of the same properties of observed traffic and it is formed in a way which appears consistent with the way Internet users generate traffic. A number of recent studies [5, 13, 21, 30] have shown that many sources of Long Range Dependent (LRD) traffic streams supply a significant part of the traffic carried on broadband networks. In [30], it was shown that one possible source of this burstiness was in the aggregation of independent on-off sources with heavy tailed, on and/or off, time distributions. In [14], it was shown that a model such as the PPBP could be considered a limiting case for the multiplexing of a large number of such independent heavy-tailed on-off sources. Thus, the PPBP is a natural candidate for the modeling of LRD packet data traffic streams.

In this section, we will present mathematical definitions of the PPBP. We note that when the Pareto-distributed burst lengths have infinite variance, the PPBP is LRD. We observe that the probability of very long bursts existing within an LRD PPBP is significant, and explain how separating the long bursts from the rest of the process can aid in PPBP queues.

2.1 Definition and Statistics

Let us denote by $\mathbf{Z}^+$ the set of non-negative integers, $\mathbf{R}$ the real numbers, and $\mathbf{R}^+$ the non-negative real numbers. We consider a continuous time process $\{B_t : B_t \in \mathbf{Z}^+, t \geq 0\}$ which represents the number of active bursts contributing work to the traffic stream at time t. We define a series of arrival times $\{\alpha_i : \alpha_i \in \mathbf{R}, i = 0,1,2,\ldots\}$ and a series of departure times $\{\omega_i : \omega_i \in \mathbf{R}, i = 0,1,2,\ldots\}$. The value of B_t increases by one at time $t = \alpha_i$ and decreases by one at time $t = \omega_i$. We define $\omega_i = \alpha_i + d_i$ and label d_i ($d_i \in \mathbf{R}^+$) the duration of the ith burst. We assume $\{\alpha_i\}$ is a non-decreasing series, i.e. $\alpha_i \leq \alpha_{i+1}$ for $i = 0,1,2,\ldots$, but we do not restrict d_i (apart from the requirement that the burst duration is positive) and so $\{\omega_i\}$ is not ordered. The value of B_t is given by

$$B_t = \sum_{i=0}^{\infty} 1_{t \in [\alpha_i, \omega_i]}$$

where

$$1_X = \begin{cases} 1, & \text{if } X \text{ is True,} \\ 0, & \text{otherwise.} \end{cases}$$

The arrival of bursts is a Poisson process with rate λ, so the intervals between adjacent burst arrival times, $\alpha_i - \alpha_{(i-1)}$, are negative exponentially distributed with mean $1/\lambda$, and the mean number of new bursts arriving in a time interval of length T is Poisson distributed with mean λT. It is well known that if the bursts arrivals are a Poisson process, the value of B_t is Poisson-distributed, with mean λ times the mean burst duration (e.g., [6]).

In the PPBP, the burst durations, d_i, are independent and identically distributed Pareto random variables, having the same distribution as random variable d. Using Pareto distributed burst durations allows the significant long bursts that characterize LRD traffic to occur in the model. The complementary distribution function of d is

$$\Pr\{d > x\} = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, & x \geq \delta, \\ 1, & \text{otherwise,} \end{cases} \tag{1}$$

$\delta > 0$. For $1 < \gamma < 2$, we have that

$$\mathrm{E}(d) = \frac{\delta\gamma}{(\gamma - 1)} \tag{2}$$

and the variance of d is infinite.

In order that the burst process should be stationary, the system is initialized with b_0 initial bursts, where b_0 is a Poisson random variable with mean $\lambda \mathrm{E}(d)$. The durations of these bursts are independent and identically distributed random variables. Their common distribution is the same as a random variable ω which is the forward recurrence time of the Pareto distribution. Thus $\alpha_i = 0$ for $i \in \{1, \ldots, b_0\}$ and ω_i values for $i \in \{1, \ldots, b_0\}$ are drawn from

$$\Pr\{\omega > x\} = \begin{cases} \frac{1}{\gamma}\left(\frac{x}{\delta}\right)^{1-\gamma}, & x \geq \delta, \\ \frac{\gamma-1}{\gamma}\left(1-\frac{x}{\delta}\right)+\frac{1}{\gamma}, & \text{otherwise.} \end{cases} \tag{3}$$

We then consider a related process, $\hat{A}_t$, the continuous time process representing the total amount of work contributed by all bursts in the period $(0,t]$. We consider the case where all bursts contribute work at a constant rate r. Thus

$$\hat{A}_t = r \int_0^t B_s ds.$$

This gives a mean of

$$\mathrm{E}(\hat{A}_t) = \frac{\lambda t r \delta \gamma}{(\gamma - 1)}.$$

Cases in which the bursts do not all contribute equal rate, or in which the work rate from a given burst may vary as a function of time, are not considered here. Results regarding the properties of burst processes in which r is not necessarily constant, or the same for all bursts, are presented in [28].

By [22], the variance of $\hat{A}_t$ can be obtained by repeatedly integrating the complementary distribution function of the burst distribution:

$$\mathrm{Var}[\hat{A}_t] = 2\lambda r^2 \int_0^t dt \int_0^u du \int_v^\infty dx \Pr(d > x).$$

This calculation is performed in [18] to give

$$\mathrm{Var}[\hat{A}_t] = \begin{cases} 2r^2\lambda t^2 \left(\frac{\delta\gamma}{2(\gamma-1)} - \frac{t}{6}\right), & 0 \leq t \leq \delta, \\ 2r^2\lambda \left\{ \frac{\delta^3\gamma}{6(3-\gamma)} - \frac{\delta^2 t\gamma}{2(2-\gamma)} - \frac{t^{3-\gamma}\delta^\gamma}{(1-\gamma)(2-\gamma)(3-\gamma)} \right\}, & t > \delta. \end{cases} \tag{4}$$

Examining the expression for the variance given in (4), we see that for large t, the dominant term is $2r^2\lambda \frac{\delta^\gamma t^{3-\gamma}}{(1-\gamma)(2-\gamma)(3-\gamma)}$. If we define $H = (3-\gamma)/2$, then we can observe that for increasing t the growth of this function is proportional to t^{2H}. This implies that for $1 < \gamma < 2$ the PPBP is *asymptotically self similar* with Hurst parameter

$$H = \frac{3-\gamma}{2}. \tag{5}$$

The conditions under which M/G/∞ processes are self-similar are examined in more depth in [27].

Note that in simulations we will consider a discrete time version of $\hat{A}_t$, where time is divided into fixed length intervals called time-slots. We choose an arbitrary value, τ, to be our time-slot size and define our discrete time process to be

$$A_j = \hat{A}_{(j+1)\tau} - \hat{A}_{j\tau} = r \int_{j\tau}^{(j+1)\tau} B_s ds. \tag{6}$$

The time-slot size, τ, may take on any value, but our usual choice is $\tau = 1$. We will use $\mu = \mathrm{E}(A_j)$ and $\sigma^2 = \mathrm{Var}(A_j)$ to denote the statistics of this discrete time process.

This discrete time process differs slightly from the processes considered in [11, 20], and also from the processes analyzed in [26, 28, 29], in that the processes considered in those works sample the value of B_t, not the value of

$\hat{A}_t$ as we do. Samples drawn from B_t can take on only discrete values, while our process is a continuous-valued, discrete-time process. Notice that if a burst starts in the middle of a time-slot and continues beyond the end of that time-slot, its contribution to the work arriving in that time-slot is $\tau r/2$, which is not necessarily integer. In limiting cases for low λ and/or high $\mathrm{E}(d)$ our process will behave in a very similar fashion to the discrete-valued processes of [11, 14, 20, 26, 28, 29].

In this paper, we discuss the queueing performance of the PPBP, and of other related processes. The queueing performance of a process is generally expressed as the queue length distribution of an infinite buffer SSQ fed by that process. However, in Section 6, and in Subsection 4.2, we shall be considering more realistic dimensioning rules, involving finite buffer SSQs (mostly zero buffer SSQs). For these finite buffer queues, the queueing performance is calculated in terms of *time congestion* values. The time congestion value for a given (finite) buffer size corresponds to the proportion of time for which the buffer is full. To avoid confusion between finite and infinite buffer scenarios, in the remainder of the paper we shall refer to time congestion values in the context of finite buffer queues, and to queue length distributions for infinite buffer queues. Both measures can provide approximations to the probability of work being lost in a finite buffer, which is the value of practical concern in network dimensioning.

2.2 Infinite Variances and Long Bursts

For $x \geq \delta$, Equation (1) shows that the Pareto distribution has a complementary distribution function which decays geometrically in the form of $x^{-\gamma}$. For $1 < \gamma < 2$, a Pareto distributed random variable will have infinite variance and finite mean. This range of values of γ is of interest to us, as for $1 < \gamma < 2$ the resulting PPBP will be LRD.

When initializing the PPBP we consider a number of initial bursts. The duration of each of these bursts has the forward recurrence time distribution given in Equation (3). Its tail decays in the form of $x^{1-\gamma}$, i.e. considerably more slowly than the corresponding Pareto distribution given in Equation (1). Notice that, for $1 < \gamma < 2$, the Pareto forward recurrence time will have infinite mean as well as infinite variance.

To illustrate the effect of infinite variances on a distribution, we compare complementary distribution functions for three different distributions in Table 1. The exponential distribution considered has a finite mean of 3 and also a finite variance. The Pareto distribution has parameters $\delta = 1$ and $\gamma = 1.5$, giving a mean of 3 once more. Unlike the exponential distribution, the Pareto distribution has infinite variance. The distribution of Pareto forward recurrence times has infinite variance *and* an infinite mean. The values given for the Pareto forward recurrence distribution are for the same value of δ and γ used to give the Pareto tail probabilities.

The sample values in the table clearly illustrate the effects of using a heavy-tailed distribution. The tail probabilities of the exponential distribution drop to be virtually zero for $x \geq 1000$. By comparison, the probability of a Pareto sample exceeding 1000 is still quite significant. The probability of a sample from the Pareto forward recurrence time distribution exceeding 1000 is even larger.

In Table 1 we can see that, even for quite moderate values of γ (we chose $\gamma = 1.5$ which gives $H = 0.75$ in the corresponding PPBP), the probability of an initial burst in the PPBP having extremely long duration is significant. The probability of an equally long burst arriving later is significantly smaller than the probability of an initial burst having that length, but is still much higher than it would be for an exponential distribution.

2.3 Long Bursts and Short Bursts Components of the PPBP

In order to separate the PPBP into long bursts and short bursts components, we consider a finite interval of length W. If we consider the PPBP over any such interval, i.e., $[t, t+W]$, for arbitrary t, then there is always a probability that some of the initial bursts will last for the entire time period. We label any such bursts as *long bursts*. All other bursts are called *short bursts*. The short bursts include: (1) those bursts that start at or before t and end before $t+W$, (2) those bursts that start after t and finish at or after $t+W$ and (3) those bursts that start after t and finish at

	$\Pr\{X > x\}$		
x	Exponential	Pareto	Pareto forward recurrence
10	0.03567	0.03162	0.2108
100	3.34×10^{-15}	0.001	0.06667
1000	0	3.162×10^{-5}	0.02108
10^4	0	1.000×10^{-6}	0.006667
10^5	0	3.162×10^{-8}	0.002108
10^6	0	1.000×10^{-9}	0.000667
10^7	0	3.162×10^{-11}	0.000211
10^8	0	1.000×10^{-12}	6.67×10^{-5}

Table 1: Comparison of sample tail probabilities

or before $t+W$. Considering these long and short bursts, we will divide the PPBP into two independent processes: (1) the *long bursts process* and (2) the *short bursts process*. The long bursts process is a stationary but non-ergodic process containing only the long bursts. The short bursts process contains all the remaining bursts, which we have labeled short bursts above.

Having made this division between the long bursts process and the short bursts process, the behaviour of the PPBP within a finite time can be more readily understood. The long bursts process will simply be a CBR component for the duration of the time period. The rate of the long bursts process will be a Poisson distributed value.

The first and second order statistics of this short bursts process over the interval $[t,t+W]$ can be shown to be stationary because:

(a) The long bursts and short bursts processes are independent; this is clear because each is formed by thinning a Poisson process, and each burst is in one process or the other, never in both.

(b) The whole process is stationary over the interval and the long bursts process is stationary over the interval, hence stationarity of the mean and covariance of the short bursts process over the interval $[t,t+W]$ follows by subtraction.

For example, if the short bursts process is denoted by S_t, the long bursts process by L_t and the total process, which is a conventional PPBP, by B_t,

$$\mathrm{Cov}(S_s,S_t) = \mathrm{Cov}(B_s,B_t) - \mathrm{Cov}(L_s,L_t).$$

Since $\mathrm{Cov}(B_s,B_t)$ depends only on $t-s$, and the same is true of L_t, this is also true of S_t.

Note that the term *stationary* here applies only to the finite time-interval processes on the time interval $[t,t+W]$. The meaning of such a term is directly analogous to the usual meaning, but expectations and events to which the stationarity condition (i.e. time invariance) applies are all constrained to lie entirely inside the interval $[t,t+W]$.

Strict stationarity of the process S in the interval $[t,t+W]$ can be shown by using basically the same argument just used to show that the autocovariance is time-invariant to show that the *moment generating functional* of the stochastic process S is time-invariant. This is proved in Appendix A.

This is not the only way to subdivide the process into long and short bursts. For example, we could subdivide the process defined on the entire real line by removing all bursts longer than W. This would be a *different* subdivision of the process into short and long bursts. Both decompositions are valid. The latter decomposition is more complicated because in the interval $[t,t+W]$ there will be long bursts which finish in the interval and ones which start in the interval.

The effect of long bursts can be seen most clearly in simulation, where we examine the PPBP over a finite time period. Regardless of the duration of the simulation, there will always be a significant probability of one or more long bursts being present. As these long bursts can have a significant impact on the properties of the process, and in particular on the queueing performance of the process, it is important that they are dealt with. This issue is the main theme of Section 3.

The same division between the long bursts process and the short bursts process is used in the development of the quasi-stationary approximation described in Section 4.1. In the quasi-stationary approximation, we separate the PPBP into long bursts and short bursts components. Existing techniques for estimating the queueing performance of stationary processes are used to estimate the performance of the short bursts process. These results are then combined according to the probabilities of the long bursts process being in given states to give an overall performance estimate for the PPBP queue.

3 Simulation Techniques

We consider a discrete time, infinite buffer SSQ with constant service rate C fed by a PPBP. Let the amount of work buffered in the queue at the end of time interval n be Q_n. In this section, we will consider simulation techniques that could be used to generate an estimate of the stationary queue length complementary distribution, $\Pr\{Q > x\}$, for any $x \geq 0$. In particular, we consider the impact of long bursts on the accuracy of simulation results for the PPBP SSQ. In Subsection 3.3 we show how to factor in the effects of the long bursts in queueing simulations in order to produce reliable results.

3.1 Long Bursts in PPBP Simulations

There is always a probability that any finite length simulation of the PPBP process should include bursts which last for the entire duration of the simulation. Depending on the parameters of the PPBP process, the probability of such long bursts occurring may be relatively small, but their effect may be quite significant. We therefore search for a method to account for the effects of such long bursts on PPBP queueing results.

We consider the case where the duration of the simulation is some time T. We will separate the PPBP into long bursts and short bursts components based on the simulation duration T, making T the simulation counterpart of the interval length W which in this paper we use for the analytical approach. The long bursts will be those which are present for the entire duration of the simulation. This means that those bursts must be among of those present at $t = 0$. By (3), for $T > \delta$, the probability of any given initial burst lasting the entirety of the simulation is

$$p = \Pr\{\omega > T\} = \frac{1}{\gamma}\left(\frac{T}{\delta}\right)^{1-\gamma}. \tag{7}$$

The total number of (long and short) initial bursts is a Poisson random variable B_0 with mean $\lambda \mathrm{E}(d)$. Let N be the number of initial bursts with duration T or greater (long bursts). Because each of the bursts in B_0 is included in N independently with probability p, N also has a Poisson distribution [23] with parameter

$$\mu_N = \lambda \mathrm{E}(d) p. \tag{8}$$

Substituting (2) and (7) into (8) gives

$$\mu_N = \frac{\lambda\delta\gamma}{\gamma-1}\frac{1}{\gamma}\left(\frac{T}{\delta}\right)^{1-\gamma} = \frac{\lambda\delta^{\gamma}}{\gamma-1}T^{1-\gamma}. \tag{9}$$

We are dividing the PPBP into two independent sub-processes. The mean of the PPBP is $\lambda \mathrm{E}(d)$. The mean of the long bursts process is μ_N given by Equation (9) above. At any point in time, the number of bursts in progress in the short bursts process is a Poisson random variable with mean $\lambda \mathrm{E}(d)(1-p)$.

3.2 Determining the Minimum Simulation Duration

For a given realization of the PPBP over a finite period of time, the long bursts process will behave as a CBR component. In an SSQ queueing simulation the effect of the long bursts process will be a reduction in the capacity available to the short bursts process. Given the duration of the simulation, and the properties of the PPBP, we can determine the probability with which the short bursts process will have a given capacity available to it. The probability of the long bursts process having rate nr for the duration of the simulation is Poisson with mean μ_N given by (9).

There is a non-zero probability that the long bursts process will reduce the capacity available to the short bursts process so much that the mean rate of the short bursts process will exceed the capacity available to it, pushing the system into an unstable state. Assuming $C > \lambda r \mathrm{E}(d)$, this instability would only be "temporary". In order to guarantee the reliability of our results, we need to ensure that the probability of an unstable simulation occurring is less than some value ε.

In an infinite buffer system the busy period caused by this instability could conceivably continue for a considerable time after the instability ceases. However, we can put a strict bound on the duration of this period of time as follows.

Consider an aggregated input process formed by collecting the work arriving in the intervals $(0,T)$, $(T,2T)$, $\ldots$, and then subtracting from each number in this sequence the work which can be completed in an interval of duration T. Let us denote this process by $\{\mathcal{W}_{kT}\}$, $k = 0, 1, \ldots$. What we have here is a very broad aggregate overview of the process, in which each number in the sequence $\{\mathcal{W}_{kT}\}$ is a summary of quite a long period of time. The great majority of entries must be negative (since the process as a whole is stable, and T has been chosen long enough that the probability that $W_{kT} > 0$ is quite small, e.g. 10^{-9}.

We seek a bound on the proportion of the intervals in the sequence $(0,T)$, $(T,2T)$, $\ldots$, during which the original process is either (i) unstable, i.e. the number of long bursts is sufficient that short bursts process is unstable in this interval, or (ii) affected by a preceding interval or sequence of intervals in which the short bursts process was unstable to the extent that the buffer is still non-empty.

Actually, we have a very good estimate and upper bound for the proportion of intervals $(0,T)$, $(T,2T)$, $\ldots$, which are unstable, namely $P\{\mathcal{W}_1 > 0\}$, which is basically, by construction, ε. The difficult issue is to know how many *other* intervals are affected by the heavy queues which build up in the buffers during overloads caused by these unstable intervals. However, this proportion can be estimated, and bounded, by the probability that the buffer in a certain queueing system is non-empty. The queueing system in question is the one where the net input process is $\{\mathcal{W}_{kT}\}$, $k = 0, 1, \ldots$. Let $\{\mathcal{V}\}_{kT}$ denote the buffer process associated with this net input process, so $\mathcal{V}_0 = 0$ and

$$\mathcal{V}_{kT} = \max(\mathcal{V}_{(k-1)T} + \mathcal{W}_{kT}), \tag{10}$$

$k > 0$. From the proposition in Section 2.4 of [1], $\Pr\{\mathcal{V}_\infty > 0\}$ is bounded by $\overline{\Phi}_\sigma\left(\Psi^{-1}_{c,\sigma}(0)\right)$ in which $\overline{\Phi}_\sigma$ denotes the Gaussian complementary distribution function with mean zero and standard deviation σ and $\Psi^{-1}_{c,\sigma}(0)$ denotes the inverse of the *normal loss function*, i.e.

$$\Psi_{c,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_x^\infty (y-c) e^{\frac{-y^2}{2\sigma^2}} dy = \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} - c\overline{\Phi}_\sigma(x),$$

evaluated at zero. The parameters $-c$ and σ here are the mean and standard deviation of $\mathcal{W}_{kT}$, and for any choice

of these parameters of real interest,

$$\overline{\Phi}_{\sigma}\left(\Psi_{c,\sigma}^{-1}(0)\right) < 3\Pr\{\mathcal{W}_{kT}' > 0\} = 3\varepsilon \tag{11}$$

(actually, $\overline{\Phi}_{\sigma}\left(\Psi_{c,\sigma}^{-1}(0)\right) \approx 2\Pr\{\mathcal{W}_{kT}' > 0\}$, but a bound is what is wanted here, so we must use the larger estimate).

Now, if we modified the original queueing system by supplying no work at all to the buffer during intervals when, originally, it was unstable, the effect would be largely limited to those periods where the aggregate queue, as defined in (10), was busy. This is still a relatively small proportion of all intervals because we have chosen ε as quite a small number. It is also possible that an interval immediately following a busy period could be affected, and such intervals are the *only* other intervals which might be affected by our modification of the process, because all other intervals will be separated from our modifications by idle periods. It is to be expected that these unstable intervals will usually occur in long clumps separated by very long periods of time, however, in the worst case, from the present point of view, it is possible that each busy period was of length 1, in which case the *extra* interval would double the number of intervals affected. Hence, the modified queueing system, in which instability never occurs, is identical to the original system except in a collection of intervals, the proportion of which is less than $2 \times 3 \times \varepsilon$.

We therefore limit the scope of this paper to systems where the probability of instability, ε, is negligible – $\frac{1}{6}\times$ the degree of accuracy sought for the distribution function in the simulation output. We know that the state space includes some states where the system is unstable, but the total probability of this part of the state space is less than ε. The flow-on effects from these states to other states still only affects a proportion 6ε of the state space. Therefore, if we alter the behaviour of our system in this region of the state space, by ignoring the possibility that we reach an unstable state, the error in our estimates of any probability at all in the system, caused by this adjustment, will be less than 6ε.

This scope limitation does not significantly weaken the results of this paper, because analytical results for heavy traffic cases are already available [25]. Our work in fact extends the scope to other, more realistic, cases.

Let ϕ be the largest value of n for which the capacity available to the short bursts process exceeds the mean rate of that part of the process. We therefore consider $\phi = \lfloor C/r - \lambda \mathrm{E}(d)(1-p) \rfloor$, where $p = \Pr\{\omega > T\}$ as in the previous subsection. If $n > \phi$ then the system will be unstable for the length of the simulation.

Increasing the duration of the simulation will reduce the probability of long bursts occurring, so we need to find the simulation time duration T which ensures that $\Pr\{N > \phi\} \leq \varepsilon$. We solve

$$\Pr\{N > \phi\} = \sum_{n=\phi}^{\infty} \frac{(\mu_N(T))^n}{n!} e^{-\mu_N(T)} = \varepsilon \tag{12}$$

to find the value of T for given values of ϕ and ε. Notice that T obtained by (12) may be too long for practical purposes. The scope of this paper is restricted to cases whereby the simulation length is practical.

We now present several sample values to give the reader an idea about the length of simulations required in order to ensure a small probability, $\varepsilon = 10^{-9}$, of the system being unstable for the duration of the simulation. We consider a family of PPBPs, all with $\sigma^2 = 0.10666\overline{6}$ and $H = 0.75$, each fed into a SSQ with capacity such that the service margin is $C - \mu = 1$. We take $\gamma = 1.5$ and $\delta = 1$ in all cases. The parameter λ is varied, and for each value of λ, (4) is solved to give the value of r. Using the above discussed methods, we determine the length of a simulation, such that the simulation is not overwhelmed by a collection of very long initial bursts.

The results are presented in Table 2. We see that the size of T decreases as λ increases. For all the cases considered, $T = 10^6$ is easily sufficient for us to be assured that the long bursts are extremely unlikely to push the system into instability.

λ	T
0.5	302578
1	2314
10	52.49
100	14.48
1 000	9.819

Table 2: Sample values of the required minimum simulation duration to avoid a significant probability of long bursts dominating behaviour

3.3 Factoring in Long Bursts

Even when the effects of the long bursts are very unlikely to cause instability, they may have a significant impact on queueing performance. To calculate simulation estimates which account for the long bursts, we simulate to calculate estimates of the queue length distribution in SSQs with service rate nr lower than the service rate of interest. The short bursts process fed into an SSQ with service rate $C - nr$ simulates the effect of a PPBP with n bursts enduring for the length of the simulation being fed into an SSQ with service rate C.

We choose the simulation time duration, T, according to (12) so as to avoid any significant impact from long bursts driving the system into instability. We consider the discrete time version of the PPBP given by (6) to be input to an infinite buffer SSQ with constant service rate C. At the end of a time-slot j, the amount of work in the buffer will be given by $Q_j = (Q_{j-1} + A_j - C)^+$, where $X^+ = \max(X, 0)$. We set $Q_0 = 0$.

To calculate the overall performance of the PPBP SSQ, we subdivide the PPBP into its long bursts and short bursts components. The long bursts and short bursts processes are as presented in Subsection 2.3 above. The time period of interest is the simulation time duration, T. All initial bursts which last the entire length of the simulation are assigned to the long bursts process. All other bursts form part of the short bursts process. This means that the short bursts process includes: those initial bursts that end before $t = T$; those bursts that start after $t = 0$ and finish after $t = T$; and those bursts that start after $t = 0$ and finish before $t = T$. The effect of the long bursts process is simulated by adjusting the service rate. To estimate the queueing performance of the short bursts process, we simulate an SSQ with service rate $C - nr$ fed by the short bursts process.

The short bursts component of the PPBP is initialized with b_0^* initial bursts, where b_0^* is Poisson with parameter $\lambda \mathrm{E}(d)(1-p)$. In the short bursts process, these initial bursts must be short bursts, so we truncate the Pareto forward recurrence time, to ensure that none of the initial bursts have length larger than T. After this initial point, the short bursts component is allowed to evolve as if it were the full PPBP.

The estimate of the performance of the short bursts process will be given by averaging the results of M simulations, each of which will include T sample intervals. In each simulation we estimate $\Pr\{Q > x\}$ by the number of time intervals in which the amount of buffered work exceeds x, divided by the total number of intervals in T. We will denote the estimate of the queue length complementary distribution given by such a set of simulations as $S^*(x; C - nr, M, T)$. Assuming that the queue size distribution process is both stationary and ergodic, this simulation estimate will approximate the queue length complementary distribution for the infinite buffer SSQ with service rate $C - nr$ fed by the short bursts process. These values of $S^*(x; (C - nr), M, T)$ are estimates of the queue length complementary distribution for the original queue (with service rate C) fed by a PPBP in which n bursts are active for the duration of the simulation.

To estimate the total queue length distribution, taking into account the probability of existing long bursts, we sum these estimates, weighted by the probability of n long bursts appearing at the start of the simulation. The

estimate of the queue length complementary distribution for the PPBP is then given by

$$\Pr\{Q > x\} = \sum_{n=0}^{\infty} \Pr\{N = n\} S^*(x; (C - nr), M, T). \tag{13}$$

Equation (13) uses the idea known as the quasi-stationary approximation. It relies on the assumption that the input process transits slowly between certain states, so slowly that the stationary behaviour of the queue is approximately the weighted average of the different behaviours in the different states, weighted by the probability of being in each state. The idea of the quasi-stationary approximation is used again in Section 4 to produce a very accurate approximation for the performance of the PPBP SSQ.

3.4 Quantum Simulation

Another related simulation technique has also been applied to this problem: quantum simulation. Quantum simulation is a method of rare event simulation in which multiple simulations evolve according to a variety of models, not necessarily independent, and not necessarily consistent individually with the original model. This aggregate of simulations is related to the original model under consideration in such a way that an appropriate weighted aggregate of all these simulations is consistent with the original model. See [2] for further details on this method and its relationship with other fast simulation techniques. Results from quantum simulations appear to be consistent with the simulation results presented in this paper, however it would add little to the present paper to include a discussion of those results here.

4 Analytical Queueing Results for the PPBP

By far the most common approach to developing approximations for stationary queueing distributions is to develop a formula which is exact for a limiting case, as a certain parameter tends to a specific value. Typically the approximation is assumed to provide a satisfactory approximation for values of this parameter which are sufficiently close to the limiting value.

For example, we might find an approximation which works well as $x \to \infty$, where x is the buffer contents, or, in the present instance, we might develop an approximation which is better and better as $\lambda \to \infty$. In fact, several such approximations are available for the stationary queueing distribution of the PPBP. The papers [9, 20, 29] provide an approximation of the first form for this process, and [1] provides an approximation of the second form.

However, with all such approximations there is the possibility that, although the approximation is good *in the limit*, it is not at all satisfactory for values of real interest, which happen to be not sufficiently close to the limiting value for the quality of the approximation to be satisfactory. In fact, this appears to be the case for both the approximations just mentioned. The approximations which hold for sufficiently large buffer values are actually quite poor for buffer values of practical use. Similarly, although there are some cases of interest where the Gaussian approximation is satisfactory, cases where λ is not sufficiently large for the Gaussian approximation to be useful are also of interest.

We present, in this section, a new approximation for PPBP SSQ performance evaluation based on the quasi-stationary idea. This approximation, which we call the *quasi-stationary approximation*, is most similar to the Gaussian approximation and is consistent with it in the sense that for sufficiently large λ, the two become closer and closer. However, the quasi-stationary approximation is better than the Gaussian approximation for small values of λ.

In this section, we present some of the analytical expressions that can be used to estimate the queueing performance of the PPBP. In the first subsection, we present the new quasi-stationary approximation. In Subsection 4.2

we give consideration to a zero buffer approximation, which could be considered an approximation which is most accurate as $x \to 0$. In Subsection 4.3 we summarize the existing results developed by Tsybakov and Georganas [29], which give an accurate approximation for $x \to \infty$. In Section 5, we will compare the estimates given by these various techniques with simulation results.

4.1 Quasi-stationary Approximation of the Queueing Behaviour of a PPBP

We will now apply the quasi-stationary idea to obtain an analytic approximation for the queue length distribution of the PPBP SSQ. The approximation is basically Equation (13), except that the simulation estimate values given by S^*, are replaced by estimates obtained by analytical means. In the analytic case, we do not have the limitation of a given simulation length, imposed by the speed computers operate, so instead of the value T, we will use the variable W, representing the minimum length of the long bursts. We will then use the quasi-stationary idea to compute an estimate for the queue length distribution as a function of W. Notice that, for each value of W, the distribution of the number of long bursts is Poisson, and the performance of the short bursts process can be obtained using known approximations for SRD queues. As in Subsection 3.3, the queue length distribution for the short bursts process given that n long bursts exist, is the queue length distribution of an SSQ fed by the PPBP excluding the long bursts, and with server rate $C - nr$.

The use of SRD queue approximations is justified, because the remainder of the process, excluding the long bursts, is made up of short bursts, and hence can be modeled as an SRD process. Each of these short bursts processes switches slowly between states dictated by the number of long bursts, and hence the quasi-stationary approximation is promising.

There are various ways of modeling the queueing behaviour of the short bursts process. One way which is convenient and may perhaps be sufficiently accurate is to model this process as Gaussian. Then, the formula of [1] is applicable. We could regard this approximation as asymptotically accurate as $\lambda \to \infty$, because for larger λ the short-range dependent process becomes more and more similar to Gaussian.

We might expect this model, in which the slowly moving and quickly moving parts of the process have been separated, to be reasonably accurate for a range of different values of W, although for quite short W, the assumption that the process moves slowly between states with a certain number of bursts longer than W would be violated and also the Gaussian approximation for the short bursts process might not be satisfactory. On the other hand, for large W, the Gaussian approximation of the short bursts process might lose accuracy for a different reason, because in this case the quasi-stationary approximation becomes closer and closer to a purely Gaussian approximation of the process, which is known, experimentally, to underestimate queueing delay [18].

In order to apply this method, we need the mean and the variance-time curve of the short bursts process. The mean of the short bursts process is $m_W = \frac{r\lambda\delta^\gamma}{\gamma-1}\left(\gamma\delta^{1-\gamma} - W^{1-\gamma}\right)$. The original PPBP is a sum of two independent components: the long bursts process and the short bursts process. Therefore, the variance-time curve, $\mathrm{Var}[\hat{A}_t]$, up to time W, of the original process, given in Equation (4) is equal to the sum of the variance time curve of the long bursts process, $v_l(t)$ and the variance time curve $v_s(t)$ of the short bursts process, i.e.

$$\mathrm{Var}[\hat{A}_t] = v_l(t) + v_s(t), \qquad 0 \le t \le W.$$

Now, the long bursts process is constant over the interval $(0, W)$, where the constant rate is a multiple of a Poisson distributed random variable with mean $\frac{\lambda W^{(1-\gamma)}}{\gamma-1}$, so

$$v_l(t) = t^2 \frac{r^2\lambda W^{(1-\gamma)}}{\gamma - 1}.$$

Hence

$$v_s(t) = \mathrm{Var}[\hat{A}_t] - t^2 \frac{r^2 \lambda W^{(1-\gamma)}}{\gamma - 1}, \qquad 0 \le t \le W. \tag{14}$$

For times longer than W, the short burst process would exhibit roughly linear growth of the variance-time curve. Actually, we will not make use of the form of this variance-time curve for values of t larger than W and it is even reasonable to argue that since the short-burst process is only defined on the interval $[0,W]$, there is no meaning for $v_s(t)$ for $t > W$.

At this point, however, let us recall the formula for the delay distribution from [1] and its derivation. Suppose V denotes the contents of a buffer supplied by Gaussian traffic with mean m, served by a server with rate c, and in which the Gaussian traffic has variance time function $v(t)$. Then the stationary complementary distribution of V has the approximation

$$\Pr\{V > x\} \approx \exp\left(-\frac{(x + (c-m)t^*)^2}{2v(t^*)} \right), \tag{15}$$

where t^* is a value $t \ge 0$ which minimizes the expression $(x + (c-m)t)^2 / v(t)$. If v is differentiable at t^*, this value can be found as a positive solution of

$$2\frac{v(t^*)}{v'(t^*)} - t^* = x/(c-m). \tag{16}$$

The quantity t^* has a simple interpretation. It is the most likely duration of the period of time over which the queue builds up to the level x. Returning to our quasi-stationary approximation, the scenario we are contemplating here is that the level x in the buffer will be achieved by the following events occurring one after the other

(i) a larger than normal number, n, of long bursts (longer than W) occurs simultaneously;

(ii) during this unusual period, the short bursts also conspire to assemble an unusual amount of work over a relatively short period of time (shorter than W).

So, in the present context, V is the buffer content of our short bursts SSQ, and t^* is the period of time over which the short bursts process accumulates sufficient work that the buffer builds to the level x. This value of t^* must be less than W or the quasi-stationary approximation is not valid. On the other hand, if t^*, as found from (16), is less than W, then the variance time curve $v(t)$ which is used in this equation will be given by Equation (14) over the range of t values from 0 to W. Thus, in solving for t^*, it is not important to define $v(t)$ for values of t larger than W.

What value should we choose for W? If we choose W too large, our approximation becomes identical to the direct Gaussian approximation of the entire system, which we know provides an estimate, but is always optimistic. For small values of W, the quasi-stationary approximation is an underestimate for a different reason. The time it takes for the buffer to build up to level x will be, in such cases, longer than W, and so, any estimate based on the *assumption* that t^* is less than W will be low. In fact, t^* is chosen to maximize the probability of the queue exceeding the threshold, so values for t^* which are forced to be below this natural maximum will simply produce low probabilities.

So, the way to choose W is simply to find the value of W which maximizes the estimate of $\Pr\{Q > x\}$.

4.2 Zero Buffer Approximations (ZBAs)

If we consider a fluid flow process with known increments process fed into an SSQ with no buffering available, then the time congestion value is simply the probability that the instantaneous arrival rate will exceed the service rate. This time congestion value can be used as an estimate of the probability of a non-empty queue in an equivalent

infinite buffer SSQ. For the PPBP, the number of active bursts at time t, B_t, is defined in Section 2, and is known to be Poisson distributed with mean $\lambda E(d)$. The instantaneous rate of work arriving is rB_t, and the zero buffer time congestion value is exactly equal to the probability that the number of active bursts will exceed the capacity of the queue. For a queue with deterministic service rate, C, this is

$$\Pr\{B_t > C/r\}. \tag{17}$$

Henceforth, we call this the *Poisson ZBA*.

In simulating the PPBP, we have considered a discretization of the system into time-slots of fixed length τ. The discrete time simulations will not give time congestion values exactly equal to the Poisson values, as there is an averaging in the discretization process, as described in (6). However, for $\tau \to 0$, the zero buffer time congestion value observed in a simulation of the PPBP should be well approximated by (17).

Note also that our simulation results, including the results given in Section 5, are queue length complementary distribution values for an infinite buffer queueing system. The value of $\Pr\{Q > 0\}$ in the infinite buffer system will exceed the time congestion value for the zero buffer case. This effect of the infinite buffer will cancel out the averaging effect discussed above to some extent.

For $\lambda \to \infty$ the PPBP will converge to a Gaussian process [3]. Another simple estimate of the probability of the infinite buffer queue being non-empty is given by considering the probability that a Gaussian random variable with the same mean, μ, and standard deviation, σ, as the PPBP, exceeds the available capacity C. This approximation will henceforth be called the *Gaussian ZBA*.

4.3 Tsybakov and Georganas Bounds

Bounds for the queueing performance of M/G/∞ processes, and specifically for an M/Pareto/∞ process similar to the PPBP, are presented in [29]. In [29] estimates of both finite queue time congestion values (proportion of time intervals when loss occurs) and finite queue loss probabilities (proportion of work arriving which is lost) are presented. Only the expressions for the time congestion values are presented here.

The upper bound is given by

$$\Pr\{Q > x\} \leq \frac{\left(\lambda\gamma\delta^{\gamma}(\gamma-1)^{-\gamma}\left(\frac{C}{r}+2\right)^{\gamma-1} r^{\gamma-1}\right)^{k}}{k!} x^{(1-\gamma)k}. \tag{18}$$

and the lower bound by

$$\Pr\{Q > x\} \geq \frac{\gamma^{k}\delta^{\gamma k} r^{(\gamma-1)k}}{\gamma(\gamma-1)^{k}\left(E(d)+(1-e^{-\rho/E(d)})^{-1}-1\right)^{\gamma+k}} x^{(1-\gamma)k}. \tag{19}$$

where

$$k = 1 + \left\lfloor \frac{C}{r} - \lambda E(d) \right\rfloor, \tag{20}$$

$$\delta = \lambda E(d) - \lfloor \lambda E(d) \rfloor, \tag{21}$$

$$\Delta = \frac{C}{r} - \left\lfloor \frac{C}{r} \right\rfloor, \tag{22}$$

and the value of ρ depends upon whether $\lambda E(d) \leq 1$ as follows: if $\lambda E(d) \leq 1$,

$$\rho = \lambda E(d), \tag{23}$$

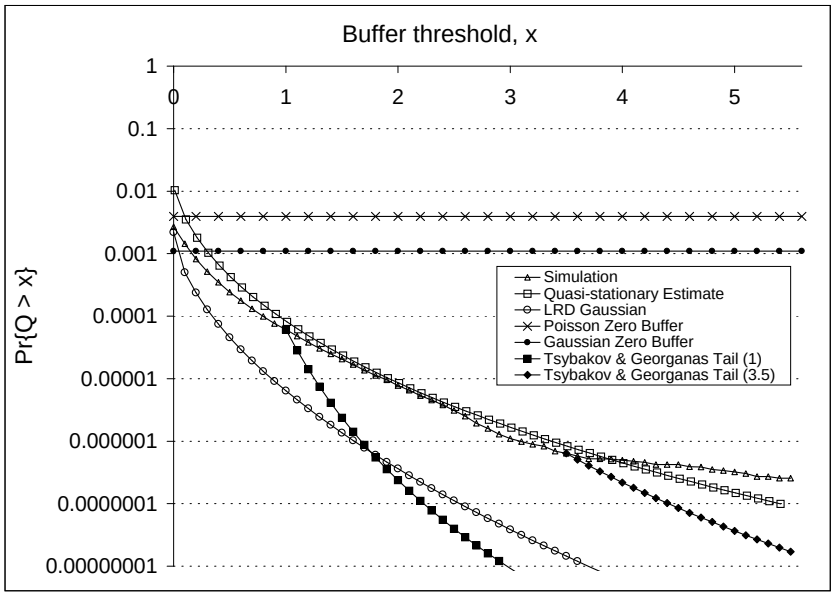

Figure 1: Comparison of the queueing estimates given for performance of a PPBP SSQ using the techniques discussed in this paper.

while if $\lambda E(d) > 1$, ρ may be any value in the range

$$0 \leq \rho < \begin{cases} 1+\delta-\Delta, & \text{if } \Delta \geq \delta, \\ \delta-\Delta, & \text{if } \Delta < \delta. \end{cases} \tag{24}$$

Both the bounds given above are valid for $x \to \infty$, and both the upper and lower bounds decay at the same rate. We would therefore predict that for large buffer sizes the PPBP process should show queueing performance in which the time congestion function decays as $x^{(1-\gamma)k}$. However, we may observe simulation results which differ from this for two reasons. Firstly, both bounds are valid only for large buffer sizes, where simulation results are likely to be least reliable, and secondly, the systems considered by Tsybakov and Georganas in deriving these bounds are not quite identical to the cases we consider. In particular, the bounds given above are for finite buffer time congestion values, while we consider the infinite buffer queue length distribution in the comparison presented below.

5 Comparison of Simulation and Analytical Results

In Figure 1, we present a typical set of results for the PPBP SSQ. We include results given by all the different techniques discussed in this paper. The specific case considered in this figure is a discrete time SSQ with service rate $C = 2.897$ units of work per interval fed by PPBP input. The PPBP fed into this queue has mean $\mu = 1.897$, variance $\sigma^2 = 0.10666$ and Hurst parameter $H = 0.75$. The parameters of the PPBP are $\lambda = 10, r = 0.0632, \delta = 1$ and $\gamma = 1.5$.

The simulation results shown in the figure are generated using the methodology discussed in Section 3.3. The quasi-stationary estimate values are calculated by the technique described in Section 4.1. We observe that the quasi-stationary estimate agrees with the simulation results.

We also compare against the LRD Gaussian estimate given in [1]. This Gaussian estimate forms the basis of the quasi-stationary estimate, but we do not divide the PPBP into long and short bursts components when calculating the LRD Gaussian estimate. The LRD Gaussian estimate is simpler to calculate than the quasi-stationary estimate, but it does not guarantee accurate results. Recall that, by [3], as $\lambda \to \infty$ the LRD Gaussian estimate will provide accurate results.

The Poisson and Gaussian ZBAs are also shown. We observe that the value of $\Pr\{Q > 0\}$ given by simulation falls between the two zero buffer estimates. Both the Poisson ZBA and the Gaussian ZBA represent reasonable estimates of the probability of the infinite buffer SSQ fed by the PPBP being non-empty.

Note that the Tsybakov and Georganas upper and lower bounds (Equations (18) and (19) respectively) differ only by expressions which are constant with respect to x. Although they are asymptotically accurate as $x \to \infty$, for the set of parameters considered here these bounds are far apart. For example, for the set of parameters considered here, (18) gives $\Pr\{Q > 1\} \leq 3.67 \times 10^{16}$ (in fact, since we are discussing a probability, $\Pr\{Q > 1\} \leq 1$ gives a tighter bound) while (19) gives $\Pr\{Q > 1\} \geq 2.06 \times 10^{-25}$. To improve the readability of the figure, we show here the rate of decay of the common tail in these bounds. The curve labelled *Tsybakov & Georganas Tail (1)* in Figure 1 shows a curve of the form $Lx^{(1-\gamma)k}$, where k is given by (20) and the weight of the tail, L is given by the value of $\Pr\{Q > 1\}$ produced by the simulation results. The curve labelled *Tsybakov & Georganas Tail (3.5)* also shows a curve of the form $Lx^{(1-\gamma)k}$, this time with the value of L matched to the value of $\Pr\{Q > 3.5\}$ produced by the simulation results. Comparing these two curves with our simulations, we see that the rate of decay of the tail matches the simulation results reasonably well as x becomes larger, however for small to moderate values of x these bounds are of limited usefulness.

In Figure 2, we present a second scenario, in which the value of λ is increased. The case considered in this figure is an SSQ with service rate $C = 7$. The PPBP fed into this queue has mean $\mu = 6$, variance $\sigma^2 = 0.10666$ and Hurst parameter $H = 0.75$. The parameters of the PPBP are $\lambda = 100, r = 0.02, \delta = 1$, and $\gamma = 1.5$. Although the level of aggregation, λ is increased over that considered in Figure 1, the service margin, $C - \mu$, and the variance and Hurst parameter of the PPBP are identical in both cases.

The six curves in Figure 2 are produced using the same techniques as described for the curves in Figure 1. We would expect the quasi-stationary estimate to continue to improve, as it is based on an assumption that the short bursts process is Gaussian, and this assumption improves as $\lambda \to \infty$, however it is not necessarily clear from these results whether such an improvement is observable. We note, however, that the quasi-stationary estimate remains a good approximation of the simulation results.

The simpler LRD Gaussian estimate is now a much better estimate of the queueing performance of the PPBP. This reflects the fact that the whole PPBP is also tending towards Gaussian as $\lambda \to \infty$.

We also note that the gap between the Poisson and Gaussian ZBAs has narrowed as λ has been increased. The value of the infinite buffer measure $\Pr\{Q > 0\}$ given by simulation continues to fall between the two zero buffer estimates, so as $\lambda \to \infty$ the Gaussian ZBA improves as a conservative estimate of probability of a non-empty buffer in the infinite buffer system.

In Figure 2, the Tsybakov and Georganas tail is shown only with a weight of L as given by the value of $\Pr\{Q > 1\}$ produced by the simulation results. As in the previous example, the actual weights for the tail, produced by Equations (18) and (19), are far apart. The rate of decay of this tail better matches that of the simulation results as x becomes larger.

6 Simple Dimensioning Rules

In this section, we consider several of the dimensioning implications of the PPBP as a model for data traffic. In particular, we highlight a simple dimensioning rule based on a Gaussian zero buffer approximation. We make use

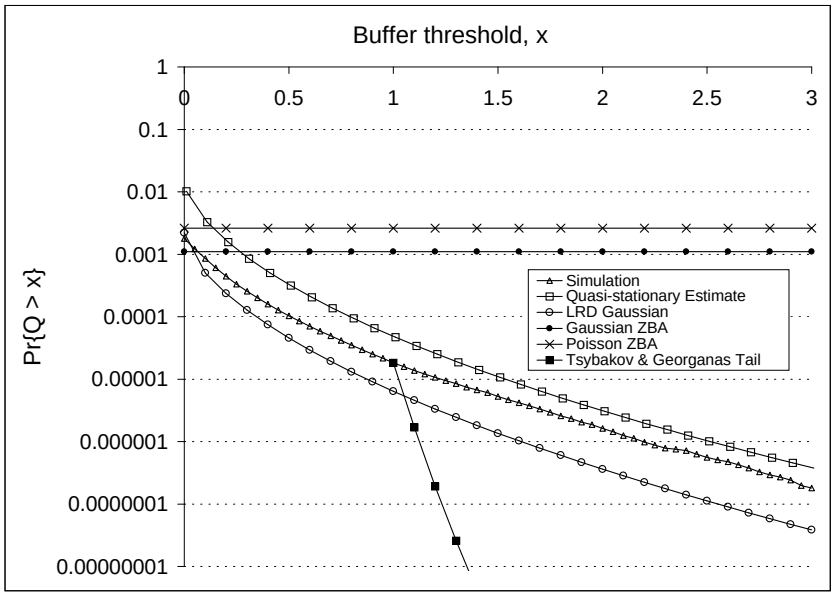

Figure 2: Comparison of the queueing estimates given for performance of a PPBP SSQ for more highly aggregated traffic.

of a PPBP fitted to real traffic to estimate the extent to which the Gaussian rule is optimistic.

6.1 The Gaussian ZBA Rule

As demonstrated in [3], as the number of independent sources being multiplexed together increases, the multiplexed stream converges towards a Gaussian process. This is true, provided that the multiplexed processes have finite mean and variance, and are of "similar" magnitude. Convergence to a Gaussian marginal distribution occurs regardless of the correlation structures of the combined traffic streams. However, the correlation structures of the combined streams *will* have an effect on the correlation structure of the aggregate stream.

Where we use a zero buffer approximation, the correlations in the stream are irrelevant, and dimensioning can be carried out based only on the properties of the marginal distribution. Any complexities in the correlation structure of the aggregate stream can then be ignored. If the traffic has a true Gaussian marginal distribution, with mean μ and variance σ^2, the time congestion value for the zero buffer case will be given by the probability that a sample from a Gaussian distribution centred at $C-\mu$ with standard deviation σ will be greater than zero.

The standard normal distribution has cumulative distribution function:

$$\Phi(x) = \Pr\{X \leq x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt \tag{25}$$

and complementary distribution function given by $\bar{\Phi}(x) = 1 - \Phi(x)$. In a zero buffer SSQ fed by a Gaussian input stream the time congestion value will be

$$\text{Time Congestion} = \bar{\Phi}\left(\frac{C-\mu}{\sigma}\right). \tag{26}$$

Note that this is the proportion of time when losses occur, which is not identical to the proportion of packets lost.

Allowable congestion, ξ	K
0.1	1.28
0.01	2.33
0.001	3.09
10^{-4}	3.72
10^{-5}	4.27
10^{-6}	4.77

Table 3: K required to give maximum congestion of ξ.

See Figure 1 of [4] for an illustration of the difference between these two probabilities in queues fed by an Ethernet trace.

If the capacity, C, is

$$C = \mu + K\sigma, \tag{27}$$

and the input traffic stream has a Gaussian marginal distribution, then the zero buffer time congestion is a function only of the value K. The choice of the parameter K will determine the probability of time congestion in the system. For a given target congestion rate of ξ, a value of K can be determined, and the capacity required for a traffic stream with known mean and variance is then given by (27). This Gaussian dimensioning rule is equivalent to Equation (4-3) in [24] and similar to the one given in [10].

The success of this dimensioning rule is dependent upon the choice of the constant K. Where the arrival process is Gaussian in nature, the required value of K can be determined simply from $\bar{\Phi}^{-1}(\xi)$. Table 3 shows some sample values of K determined in this way. For non-Gaussian processes (27) still provides a simple dimensioning rule, however using the value of K used for a Gaussian arrival process may not be appropriate. In [12] simulation is used to calculate the values of K required for real traffic streams.

The Gaussian ZBA rule gives significant scope for multiplexing gain. We consider multiplexing M identical Gaussian streams to form a single aggregate traffic stream. Each of the M streams has mean μ and variance σ^2. The aggregate stream will be a Gaussian traffic stream with mean $\mu_{tot} = M\mu$ and variance $\sigma^2_{tot} = M\sigma^2$. We assign capacity to this aggregate stream.

In the curve labeled *Gaussian ZBA* in Figure 3 we examine the efficiency gains when M identical Gaussian processes are multiplexed in a zero buffer system. The base process is Gaussian with the same mean and variance as the IP traffic trace examined in [17], i.e., $\mu = 5225$ and $\sigma^2 = 21.22 \times 10^6$. We then consider the process created by multiplexing M independent copies of this base process. The result will be a Gaussian process with mean $\mu_M = M\mu = 5224.66M$ and variance $\sigma^2_M = M\sigma^2 = 21.22 \times 10^6 M$. Using the Gaussian ZBA rule from (27), and choosing $K = 4$, we assign capacity $C = \mu_M + 4\sigma_M$ to this process. This value of K gives a time congestion value of $\xi = 3.17 \times 10^{-5}$. The efficiency is given by μ_M/C.

The second curve in Figure 3 shows the efficiency gains when capacity is assigned according to a Poisson ZBA. For $M = 1$, the process is a PPBP fitted to the same IP trace considered for the Gaussian dimensioning. We have examined this IP trace in [17]. Using repeated simulation, we have found that a PPBP with parameters $\lambda = 0.4, r = 3060, \delta = 0.9153, \gamma = 1.18$ models the IP trace, and we have found that this process can accurately predict the queueing performance of the original trace in an SSQ. A CBR component labeled κ is added to the PPBP in the fitting of the real traffic carried out in [17]. The relevant value of κ for this stream is $\kappa = -2119$.

The PPBP equivalent to the multiplexing of M independent copies of the fitted PPBP has parameters $\lambda_M = M\lambda, r_M = r, \delta_M = \delta, \gamma_M = \gamma$ and $\kappa_M = M\kappa$. Like the Gaussian process, this PPBP will have mean $\mu_M = M\mu$ and variance $\sigma^2_M = M\sigma^2$. As in the Gaussian case, we determine the capacity, C, required such that the ZBA predicts a

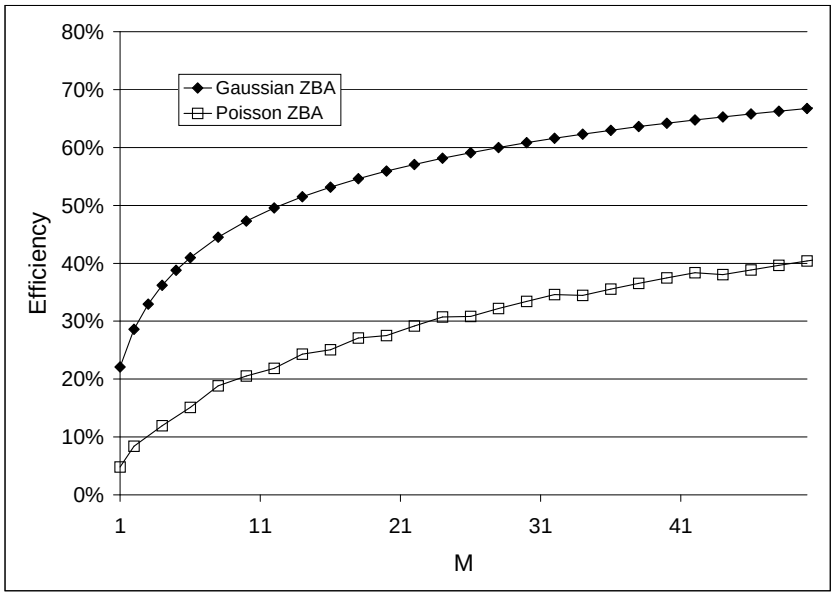

Figure 3: Improvement in efficiency with increasing multiplexing.

time congestion of $\xi = 3.17 \times 10^{-5}$ and plot the efficiency given by μ_M/C for each value of M.

We see from the figure that the Gaussian rule significantly under-estimates the capacity required by the PPBP.

6.2 Rate of Convergence to Gaussian

In the previous subsection we have shown that using the Gaussian ZBA rule gives a reasonably simple dimensioning rule. We will now examine the effectiveness of the Gaussian ZBA rule as a simple dimensioning rule for realistic streams.

We have shown in [17, 18] that real traffic streams can be modeled using the PPBP. We have also observed that if multiple PPBPs are multiplexed together, the queueing performance of the overall aggregate stream tends towards the performance of an LRD Gaussian stream [4, 18].

As the multiplexing level in the PPBP increases, we expect to see the PPBP behaving more like a Gaussian process. We will evaluate the closeness of the PPBP to Gaussian by examining the time congestion function obtained by feeding the PPBP into a finite buffer SSQ with service rate determined by the Gaussian zero buffer approximation. As the level of multiplexing increases, we would expect to see the accuracy of this approximation improving, i.e., the time congestion value in the SSQ fed by the multiplexed PPBP should approach the allowable time congestion used for dimensioning.

The multiplexing of realistic traffic streams is simulated by creating scaled PPBPs. Once again, we consider a PPBP fitted to the IP traffic byte stream examined in [17]. This stream contains measurements of the number of bytes arriving on a single link measured in 0.1 second intervals. The measurement was made in early 1999. The stream has a mean rate of $\mu = 5225$, variance of $\sigma^2 = 2.122 \times 10^7$ and a Hurst parameter of $H = 0.91$. We have found that a PPBP with parameters: $\lambda = 0.4, r = 3060, \delta = 0.9153, \gamma = 1.18$ and $\kappa = -2119$; accurately predicts the behaviour of the real traffic.

We then use the PPBP to model the effect of increasing levels of multiplexing in the IP traffic. As before, the PPBP equivalent to the multiplexing of M independent copies of the original PPBP has parameters $\lambda_M = M\lambda, r_M =$

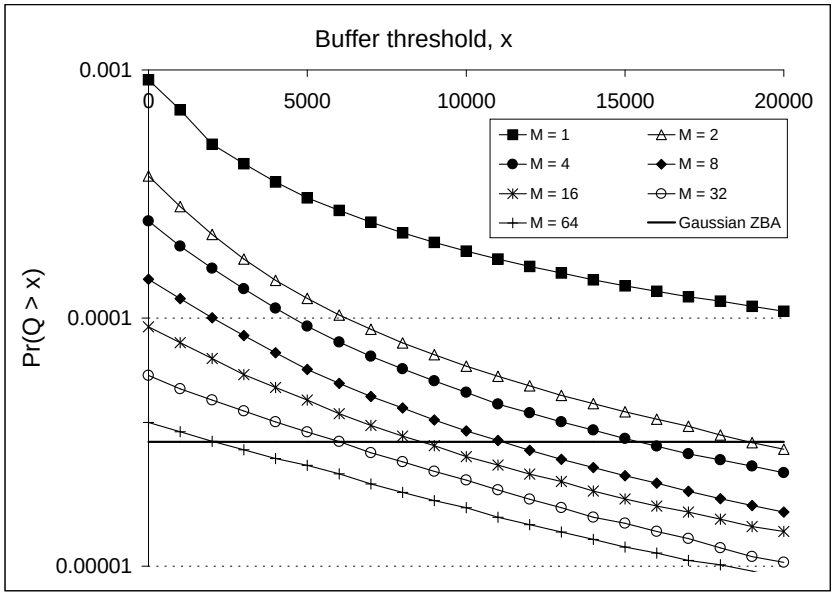

Figure 4: Comparison of PPBP time congestion and time congestion value predicted by Gaussian ZBA.

M	Threshold for Gaussian ZBA congestion	Delay (ms)
1	74 150	313
2	18 900	51.8
4	15 450	26.7
8	11 100	11.8
16	8 650	5.50
32	6 000	2.22
64	2 100	0.43

Table 4: Buffering required to achieve allowable time congestion level.

$r, \delta_M = \delta, \gamma_M = \gamma$ and $\kappa_M = M\kappa$. This process will have mean $\mu_M = M\mu$ and variance $\sigma_M^2 = M\sigma^2$. The Hurst parameter will be unchanged by the level of multiplexing.

Figure 4 shows results for this set of PPBPs. The dimensioning rule used is $C = \mu + 4\sigma$, giving a maximum Gaussian time congestion of 3.17×10^{-5}. We see that as M increases, the zero buffer time congestion value, $\Pr\{Q > 0\}$, approaches the target value permitted under the dimensioning rule. The results shown are given by simulation. Although error bars are not shown in the figure, the zero buffer approximation falls within the 95% confidence interval of the $\Pr\{Q > 0\}$ values for $M = 64$ but not for $M = 32$.

Table 4 shows the amount of extra buffering required to compensate for the fact that real traffic is not Gaussian. This is minimum buffer size for which the time congestion value is less than the time congestion value given by the Gaussian zero buffer approximation of 3.17×10^{-5}. To give an idea of the impact of this extra buffering we also show the maximum delay which may be encountered by data arriving at the buffer.

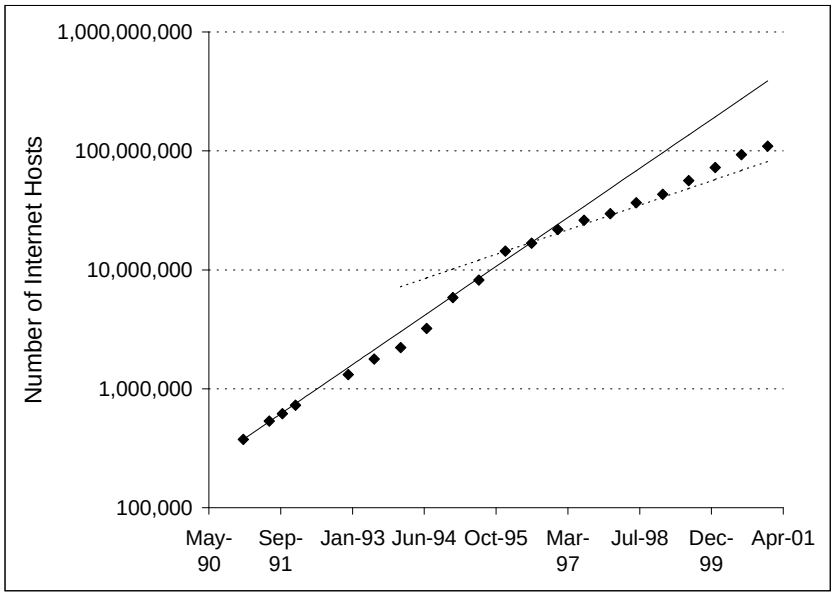

Figure 5: Growth in number of IP hosts.

6.3 When Will Convergence to Gaussian Happen?

The results given in the previous subsections indicate that the Gaussian ZBA is still a gross approximation for real data traffic on fine timescales. However, we also see that the usefulness of the Gaussian ZBA as a conservative dimensioning rule at these timescales improves considerably as the level of multiplexing increases. Here we present some of the evidence which suggests that the level of multiplexing in real networks is increasing.

Figure 5 shows the growth in the number of hosts connected to the public Internet. The diamonds show counts of the number of hosts, as recorded in the Internet Domain Survey at the Internet Software Consortium (*http://-www.isc.org/ds/*). The solid line represents a doubling in the number of hosts occurring every 12 months. This gives a good estimate of the growth of the Internet for the period between January 1991 and January 1996. The dashed line represents a doubling every 24 months, and appears to give a good estimate of the growth rate since January 1996.

More direct evidence of the increase in the amount traffic being carried across the Internet has been compiled by Odlyzko [19]. Evidence from a range of sources is summarized in [19] and indicates that the amount of traffic being carried on the Internet is doubling every twelve months.

Thus we see that, although direct measurement of the Internet is becoming increasingly difficult, the best evidence suggests that the amount of traffic carried across the Internet is growing exponentially. Further, the increasing number of Internet hosts suggests that the growth is at least partly caused by an increasing number of streams being carried, suggesting that the level of multiplexing on individual links is also increasing. The trace modeled in Section 6.2 was recorded in early 1999. Since that time the total traffic on the Internet has probably quadrupled, and the number of Internet hosts at least doubled. Assuming that this overall increase translates to similar increases on individual links, there is every reason to suspect that the level of multiplexing on the measured link is likely to have at least doubled in the period since that measurement was made.

If Internet traffic growth results primarily from an increasing number of streams (as opposed to an increase in the bit rate per stream), then a doubling in load will translate to a doubling in multiplexing level. On current trends,

the traffic on the backbone IP link where the 1999 trace was recorded will become close to Gaussian by 2004 under this assumption. Alternatively, if we assume that the level of multiplexing is directly proportional to the number of separate hosts, then the traffic on this link will be close to Gaussian by around 2010. It is likely that there are already core links with Gaussian traffic.

7 Conclusions

We have presented new insights into the behaviour of SSQs fed by the PPBP. We have developed our ideas by dividing the PPBP into two sub-processes; a long bursts process and a short bursts process, and considering the impact of each of these sub-processes separately.

We have seen that, irrespective of the duration of a simulation, the existence of long bursts at the beginning of the simulation may impact on the reliability of the simulation. We have described an improved methodology for simulation of the PPBP which takes into account the impact of long bursts. We have also developed a bound on the simulation duration required to ensure reliable results.

A quasi-stationary approximation for the performance of a PPBP SSQ, which is also based on the idea of treating long bursts separately, has also been presented. Comparisons with simulation results have shown that the quasi-stationary approximation accurately predicts the performance of a PPBP SSQ.

Simple link dimensioning and multiplexing gain evaluations based on Poisson and Gaussian ZBAs were given. We have demonstrated that the less conservative Gaussian ZBA becomes more useful as traffic becomes Gaussian.

Finally, the question of how close is traffic in today's networks to being Gaussian was addressed. By means of both simulations and the quasi-stationary approximation which was developed in this paper it is possible to see explicitly how significant is the difference between a burst model of network traffic and a Gaussian model. Apparently, in some cases, the difference is not so great. It seems likely that there are already places in today's networks where traffic can be regarded close enough to Gaussian. Of course, this depends on the level of aggregation, which will always vary from place to place.

A Proof of Strict Stationarity of the Short Burst Process

The same argument used in Subsection 2.3 to show second-order stationarity can be applied to the moment generating functionals, $\Phi_S(\theta)$, $\Phi_B(\theta)$, and $\Phi_L(\theta)$, defined, for example, by

$$\Phi_S(\theta) = E\{e^{\int_t^{t+W} S_s d\theta(s)}\}, \tag{28}$$

for θ any peicewise continuous real-valued function defined on $[t, t+W]$. (Aside from a coefficient, i, in the exponent this is the definition of the characteristic functional as given in [8]). As we now show, this will demonstrate strict stationarity of the process S.

The moment generating functional of a stochastic process, e.g. S as at (28), completely characterizes the process. To see this, observe that if we choose for θ a function which takes the value $\sum_{j=1}^{i} \theta_j$ on (t_i, t_{i+1}), $i = 0, \ldots, n$, where $t_0 = t$, $t_{n+1} = t+W$ and $t_i \in [t, t+W]$, $i = 1, \ldots, n$, then, as as function of θ_i, $i = 1, \ldots, n$, $\Phi_S(\theta_1, \ldots, \theta_n) = \Phi_S(\theta)$ is the joint moment generating function of $S_{t_1}, \ldots, S_{t_n}$. It follows that the joint distribution of any collection of values of the process S is completely determined by Φ_S and hence this stochastic process is completely characterized by its moment generating functional.

By independence of the processes S and L, and using the fact that $B = S + L$, we find $\Phi_B(\theta) = \Phi_S(\theta) \times \Phi_L(\theta)$, so

$$\Phi_S(\theta) = \Phi_B(\theta) / \Phi_L(\theta), \tag{29}$$

for θ any peicewise continuous real-valued function defined on $[t, t+W]$. Because the functions θ take real values only, the moment generating functionals have no zeros, so the division in (29) does not present a problem.

Strict stationarity of S will follow from the fact that the characteristic functional, $\Phi_S(\theta)$, is invariant under suitable time shifts of θ. In the present case, since we are considering *strict stationarity on the interval* $[t, t+W]$, by a suitable time shift we mean any time shift which does not push the support of θ outside the interval $[t, t+W]$. (The support of $\theta(s)$ is the set of s where $\theta(s) \neq 0$.) Since this is true of the process B and of the process L, because they are both strictly stationary, by (29), it is also true for S. Thus, S is also strictly stationary.

References

[1] R. Addie, P. Mannersalo, and I. Norros, "Performance formulae for queues with Gaussian input," In *Proceedings of ITC 16*, pp. 1169–1178, June 1999.

[2] R. G. Addie, "Quantum simulation: Rare event simulation by means of cloning and thinning," To appear in *Proceedings of Modsim 2001*, December 2001.

[3] R. G. Addie, "On the weak convergence of long range dependent traffic processes," *Journal of Statistical Planning and Inference*, vol. 80, pp. 155–171, 1999.

[4] R. G. Addie, M. Zukerman, and T. D. Neame, "Broadband traffic modeling: Simple solutions to hard problems," *IEEE Communications Magazine*, vol. 36, no. 8, August 1998.

[5] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range-dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 1566–1579, February/March/April 1995.

[6] D. R. Cox and V. Isham, *Point Processes*, Chapman and Hall, 1980.

[7] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, December 1997.

[8] I. M Gel'fand and N. Ya. Vilenkin. *Generalized Functions, Volume 4: Applications of Harmonic Analysis*. Academic Press, 1964.

[9] Bárbara González-Arévalo and Gennady Samoridnitsky, "Buffer content of a leaky bucket system with long-range dependent input traffic," Technical report, Cornell University, 2000.

[10] R. Guérin and L. Gün, "A unified approach to bandwidth allocation and access control in fast packet-switched networks," In *Proceedings of IEEE Infocom '92*, May 1992.

[11] M. M. Krunz and A. M. Makowski, "Modeling video traffic using M/G/∞ input processes: A compromise between Markovian and LRD models," *IEEE Journal on Selected Areas in Communications*, vol. 16 no. 5, June 1998.

[12] T. K. Lee and M. Zukerman, "Efficiency comparisons between different model-based and measurement-based connection admission control schemes under heavy traffic," In *Proceedings of Globecom '99*, 1999.

[13] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[14] N. Likhanov, B. Tsybakov, and N. D. Georganas, "Analysis of an ATM buffer with self-similar ("fractal") input traffic," In *Proceedings of IEEE Infocom '95*, 1995.

[15] M. Mandjes, "A note on queues with M/G/∞ input," *Operations Research Letters*, vol. 28, pp. 233–242, 2001.

[16] T. D. Neame, M. Zukerman, and R. G. Addie, "Applying multiplexing characterization to VBR video traffic," In *Proceedings of ITC 16*, June 1999.

[17] T. D. Neame, M. Zukerman, and R. G. Addie, "A practical approach for multimedia traffic modeling," In *Proceedings of Broadband Communications '99*, November 1999.

[18] T. D. Neame, M. Zukerman, and R. G. Addie, "Modeling broadband traffic streams," In *Proceedings of Globecom '99*, December 1999.

[19] A. Odlyzko, "The current state and likely evolution of the Internet," In *Proceedings of Globecom '99*, pp. 1869–1875, November 1999.

[20] M. Parulekar and A. M. Makowski, "Tail probabilities for M/G/1 input processes (I) preliminary asymptotices," *Queueing Systems*, vol. 27, pp. 271–296, 1997.

[21] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[22] J. Roberts, U. Mocci, and J. Virtamo, editors, *Broadband Network Teletraffic, Final Report of Action COST 242*, Springer, 1996.

[23] S. Ross, *A First Course in Probability*, Macmillan Publishing, 1976.

[24] M. Schwartz, *Broadband Integrated Networks*, Prentice-Hall, 1996.

[25] K. P. Tsoukatos and A. M. Makowski, "Heavy traffic limits associated with M/G/∞ input processes," *Queueing Systems*, vol. 34, no. (1–4), pp. 101–130, 2000.

[26] B. Tsybakov and N. D. Georganas, "On self-similarity in ATM queues: Definitions, overflow probability bound, and cell delay distribution," *IEEE/ACM Transactions on Networking*, vol. 5, no. 3, pp. 397–409, June 1997.

[27] B. Tsybakov and N. D. Georganas, "Self-similar processes in communications networks," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1713–1725, September 1998.

[28] B. Tsybakov and N. D. Georganas, "Self-similar traffic and upper bounds to buffer-overflow probability in an ATM queue," *Performance Evaluation*, vol. 32, pp. 57–80, 1998.

[29] B. Tsybakov and N. D. Georganas, "Overflow and losses in a network queue with a self-similar input," *Queueing Systems*, vol. 35, pp. 201–235, 2000.

[30] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, 1997.